RELEVANCE RANKING IN ONLINE-KATALOGEN: INFORMATIONSSTAND UND PERSPEKTIVEN

von Otto Oberhauser und Josef Labner

Zielsetzung

Bekanntlich führen Suchmaschinen wie Google & Co. bei der Auflistung der Suchergebnisse ein „Ranking" nach „Relevanz" durch, d.h. die Dokumente werden in absteigender Reihenfolge entsprechend ihrer Erfüllung von Relevanzkriterien ausgeben. In Online-Katalogen (OPACs) ist derlei noch nicht allgemein übliche Praxis, doch bietet etwa das im Österreichischen Bibliothekenverbund eingesetzte System Aleph 500 tatsächlich eine solche Ranking-Option an (die im Verbundkatalog auch implementiert ist).
Bislang liegen allerdings kaum Informationen zur Funktionsweise dieses Features, insbesondere auch im Hinblick auf eine Hilfestellung für Benutzer, vor. Daher möchten wir mit diesem Beitrag versuchen, den in unserem Verbund bestehenden Informationsstand zum Thema „Relevance Ranking" zu erweitern. Sowohl die Verwendung einer Ranking-Option in OPACs generell als auch die sich unter Aleph 500 konkret bietenden Möglichkeiten sollen im folgenden näher betrachtet werden.

Was ist Relevance Ranking ?

„Das Ranking beschreibt den Prozess, in dem die Suchmaschine die Ergebnisse strukturiert und in einer bestimmten Reihenfolge ausgibt. Die Ausgabereihenfolge variiert mit der Art des Rankings. Die „klassische" Variante des Rankings überprüft die Ergebnisse danach, wie oft die gesuchten Stichworte in der Seite vorkommen, wo auf der Seite die Stichworte vorkommen (Gewichtung in absteigender Reihenfolge: Domain oder URL, Titel, Überschrift, Meta-Information, Dokument) und wie viele Links von dieser Seite ausgehen und auf diese Seite weisen. Eine weitere Variante ist Listung gegen Geld, bei der derjenige ganz oben steht, der am meisten dafür gezahlt hat. Allerdings gibt es nicht genügend Kunden, die dafür bezahlen wollen, sodass die restlichen Ergebnisse durch die „klassische" Variante aufgefüllt werden. Die dritte Variante ist das Nutzerverhalten, bei der die Seiten oben stehen, die von den meisten Nutzern, die nach diesen Stichwörtern gesucht haben, angeklickt wurden." (Güde & Götsch, 2002, 2)

Eines der bekanntesten Beispiele ist zweifellos Googles „PageRank"-Verfahren, das die weitverzweigte Link-Struktur des WWW für die Einschätzung der Qualität bzw. Relevanz einer Web-Seite heranzieht:

„Der Kern ist dabei, dass Google einen Link von Seite A zu Seite B als ein „Votum" von Seite A für Seite B interpretiert. Aber Google sieht sich mehr als nur das Ausmaß der Zustimmung oder der Links auf einer Seite an; Google analysiert ebenfalls die Seite, die das Votum abgegeben hat. Das Votum von einer Seite, die selber „wichtig" ist, zählt mehr und hilft, andere Seiten „wichtig" zu machen. Wichtige Websites mit hoher Qualität bekommen einen höheren PageRank, den Google sich mit jeder Suchanfrage merkt. Natürlich bedeuten Ihnen wichtige Seiten nichts, wenn sie nicht Ihr Suchwort treffen. Deshalb kombiniert Google PageRank mit einer komplexen Textsuche, um Seiten zu finden, die sowohl Ihre Suchanfrage treffen als auch wichtig sind. Google geht in der Analyse weit über die bloße Anzahl der Suchbegriffe, die auf einer Seite zu finden sind, hinaus und untersucht alle inhaltlichen Aspekte (wie auch den Inhalt der Seiten, die mit dieser Seite verbinden), um zu bestimmen, ob sie ein guter Treffer für Ihre Suche ist." (Google, 2002)

Die Rangbildung bei der Ausgabe von Dokumenten wurde allerdings schon lange Zeit vor dem Zeitalter der Suchmaschinen bzw. des WWW vorgeschlagen. So heisst es etwa im klassischen Lehrbuch des Information Retrieval (Salton & McGill, 1987, 156f.):

„Am einfachsten lässt sich das Retrievalergebnis dadurch verbessern, dass die binäre Deskribierung durch eine gewichtete Deskribierung ersetzt wird, und dass die durch eine Suchanfrage nachgewiesenen Dokumente in einer Rangfolge ausgegeben werden, die der Ähnlichkeit der Dokumente mit der Suchanfrage entspricht. Die Bildung einer *Rangreihe* erhöht nicht nur die Benutzerzufriedenheit, sondern auch die Retrievalpräzision. Ferner lässt sich durch eine Rangreihe das Relevanzfeedback-Verfahren wesentlich verbessern, da die relevanten Dokumente dem Benutzer vor den irrelevanten nachgewiesen werden; diese Anordnung führt dann umgekehrt zur Generierung neuer, präziserer Suchanfragen."

Obzwar in unseren Katalogen keine gewichtete, sondern eine binäre Indexierung durchgeführt wird (d.h. ein Deskriptor wird zugeteilt oder nicht zugeteilt) und bei der Suche im OPAC lediglich Verfahren, die auf der Boole'schen Logik basieren, zur Anwendung kommen, kann dennoch eine „weighted-term search logic" eingesetzt werden. Dabei geht es um die Rangreihung der Treffer (oder auch die Begrenzung ihrer Zahl), die bei einer Suche mittels einer unter Vewendung der Boole'schen Operatoren gebildeten Suchanfrage resultieren. Die gefundenen Dokumente werden für die Ausgabe nach einem für sie ermittelten Gewicht absteigend gereiht,

wobei die Gewichte auf der Basis der Häufigkeit des Auftretens der Begriffe in den Dokumenten bzw. der inversen Häufigkeit ihres Auftretens in der Datenbank berechnet werden (ROWLEY & FARROW, 2000, 134–137).

Relevance Ranking auch in OPACs ?

Schon vor Jahren argumentierte HILDRETH (1995), dass die in der modernen Information-Retrieval-Forschung erfolgreich getesteten Gestaltungsprinzipien und Suchmethoden auch in die Benutzeroberflächen der bibliothekarischen OPACs Eingang finden sollten, da sie ihre Überlegenheit gegenüber den Boole'schen Suchverfahren unter Beweis gestellt hätten. Im einzelnen nennt er Deskriptorengewichtung, Best-Match-Searching, Relevanzfeedback und *Ranking der Treffer*.

Im Hinblick auf das zuletzt genannte Feature und auf bibliothekarische Online-Kataloge meinte etwa auch GÖDERT schon vor geraumer Zeit (1996, 80), „dass es für zukünftige Such- und Findeinstrumente nicht mehr ausreichen kann, eine Ausgabe von bibliographischen Beschreibungen – im derzeit günstigsten Fall – in alphabetisch oder chronologisch geordneten Listen anzubieten." Eine mögliche Alternative sei Relevance Ranking: „Hiermit ist gemeint, die gefundenen Treffermengen in einer systemseitig berechneten Reihenfolge hinsichtlich ihres vermuteten Relevanzgrades für die Suche auszugeben. Angesichts verschiedener Ergebnisse von Benutzerbefragungen, wird dieser Methodik allgemein eine grosse Bedeutung beigemessen." (ibid.)

Zum selben Zeitpunkt befürchtete LEPSKY (1996), dass es noch ein weiter Weg bis zum Einsatz von Relevance Ranking, Fuzzy Suche bzw. Hypertext-Links in OPACs sei, gab aber zugleich zu bedenken, dass „der Einsatz von z.B. Algorithmen zum Relevance-Ranking oder einer sog. Fuzzy-Suche (...) umgekehrt auch nur dann sinnvoll möglich [sei], wenn eine ausreichende Textbasis als Grundlage für die Rechenoperationen zur Verfügung steht." Und weiter: „Die Gestaltung fortschrittlicher Information-Retrieval-Systeme für Bibliotheken führt daher zunächst über die Veränderung und die Erweiterung des bibliothekarischen Dokuments." (beides S. 66)

Während diese Aussage wohl auf die Möglichkeiten der Anreicherung von OPACs durch automatische Indexierungsverfahren oder durch die Hinzunahme von Abstracts, Inhaltsverzeichnissen oder gar Volltexten abzielt, sollte trotzdem auch überlegt werden, inwieweit ein gewichteter Output in OPACs mit lediglich konventionellen bibliographischen Datensätzen sinnvoll ist. Dazu kann man natürlich unterschiedlicher Meinung sein; so meinte etwa ein Teilnehmer an einer 1999 in der InetBib-Liste durchgeführten OPAC-Diskussion: „... wer in Bibliothekskatalogen recherchiert, sollte

auf jeden Fall die ausdrückliche Möglichkeit haben, ohne Fuzzy-Techniken nach exakten Daten zu suchen und eine alphabetisch geordnete Liste vorzufinden." Derselbe konzedierte aber auch, dass man Ranking in kleinen OPACs zur Erhöhung der Trefferzahl (Oder-Verknüpfung der Suchbegriffe!) bzw. in grösseren OPACs zur Ergebniseinschränkung anbieten könne. Die exemplarische Skepsis dieses Teilnehmers schliesst immerhin nicht die *Möglichkeit* aus, im OPAC Relevance Ranking als *Option* anzubieten.

Eine weitere Frage wäre, ob ein nach Relevanz sortierter Trefferoutput gar die *Standardeinstellung* des OPACs oder eben nur eine vom Benutzer auswählbare Option darstellen sollte. Wie die im folgenden Abschnitt präsentierten Daten zeigen, gibt es heute tatsächlich schon Bibliotheks-OPACs, die als Default eine Reihung der Treffer nach Relevanz implementiert haben.

Bieten OPACs heute bereits Relevance Ranking an ?

Im folgenden soll durch einen „Rundblick" bei einer Auswahl von bekannten und weniger bekannten OPACs ermittelt werden, inwieweit in diesen Katalogen das Feature Relevance Ranking eingesetzt wird bzw. wie es sich ggfs. dem Benutzer darbietet.

(1) Zunächst wurde eine Reihe von OPACs, die mit dem System Aleph 500 betrieben werden, auf das Vorhandensein der Funktion „Ergebnisliste gewichten" überprüft. Sofern Relevance Ranking angeboten wird, sieht man bei der Anzeige einer Ergebnisliste einen Button „Gewichten" („Rank", „Pertinence"), nach dessen Betätigung ein weiteres Dialogfenster erscheint, in welchem der Benutzer die für die Gewichtung herangezogenen Terme bestätigen muss. Die danach resultierende Ergebnisliste ist absteigend nach den vom System errechneten Gewichten sortiert.

Aus Tabelle 1 geht hervor, dass diese Ranking-Funktionalität derzeit allerdings nur von der Minderheit der untersuchten Aleph-OPACs – grösseren wie auch kleineren – angeboten wird. Wie sich herausstellte, werden nur im Österreichischen Verbundkatalog auch die Gewichte in der Ergebnisliste angezeigt, wogegen bei den übrigen OPACs lediglich eine (nicht näher kommentierte) Umsortierung der Kurztitel erfolgt. Hilfe bzw. Erläuterungen sind meist spärlich bzw. gar nicht vorhanden (z.B. University College London, Université Paris III, Santa Barbara). Alle Kataloge wiesen beim Ranking schon bei relativ kleinen Treffermengen eine unzufriedenstellende Performance auf.

Der Vollständigkeit halber sei auch erwähnt, dass in Aleph-OPACs keine Möglichkeit besteht, die Funktion Relevance Ranking *voreinzustellen* – weder kann der Benutzer festlegen, dass seine Suche sofort zu einer nach

Relevanz sortierten Ergebnisliste führen soll, noch ist es dem Systembibliothekar möglich, Relevanz als primären Sortierschlüssel zu definieren.

OPAC	Anzahl Titelsätze	Ranked Output	Anzeige Gewichte
Österreichischer Verbundkatalog	3,700.000	**Ja**	**Ja**
Universität Wien	1,500.000	Nein	–
Universität Graz	1,400.000	Nein	–
Universität Innsbruck	900.000	Nein	–
Universität Salzburg	1,000.000	Nein	–
Verbund für Bildung und Kultur	800.000	**Ja**	Nein
Vorarlberger Landesbibliothek	300.000	**Ja**	Nein
KOBV Berlin-Brandenburg	32,200.000	Nein	–
HBZ VK-NRW	11,900.000	Nein	–
Universität Düsseldorf	1,600.000	Nein	–
Universität Essen	900.000	Nein	–
NEBIS (Schweiz)	2,000.000	Nein	–
IDS Basel/Bern	2,000.000	Nein	–
Technische Universität Delft	k.A.	Nein	–
Universität Gent	800.000	Nein	–
King‘s College, London	k.A.	Nein	–
University College London	k.A.	**Ja**	Nein
Université de Paris III	300.000	**Ja**	Nein
Harvard University	9,000.000	Nein	–
University of Maryland	k.A.	Nein	–
Notre Dame University	2,000.000	Nein	–
Univ. of California Santa Barbara	2,500.000	**Ja**	Nein
Melvyl (California)	23,000.000	Nein	–

Tabelle 1: Verfügbarkeit von Relevance Ranking in ausgewählten Aleph 500 OPACs

(2) Des weiteren wurden auch OPACs anderer Bibliothekssysteme im Hinblick auf das Feature Relevance Ranking untersucht. Zumindest in den von uns ausgewählten OPACs der Systeme BIS, BLPC, GEAC, INNOPAC, SISIS, TALIS sowie in den Katalogen des BVB und von LIBRIS konnten wir dieses Feature nicht lokalisieren, wohl jedoch in manchen/allen OPACs der Systeme CONTEC C2, PICA, SIRSI, VOYAGER sowie im britischen Verbundkatalog COPAC (Tabelle 2).

System	Name	Anzahl Titelsätze	Ranked Output
BIS, Horizon	Südwestdeutscher Bibliotheksverbund	9,600.000	Nein
BISCwit	Niederösterreichische Landesbibliothek	300.000	Nein
BISCwit	Parlamentsbibliothek	100.000	Nein
BLPC	British Library	10,000.000	Nein
Contec C2	Central Queensland Institute of TAFE	k.A.	**Ja**
Eigenentwicklung	BibliotheksVerbund Bayern	12,300.000	Nein
Eigenentwicklung	Britischer Verbundkatalog (COPAC)	20,000.000	**Ja**
Geac	Oxford University	8,000.000	Nein
Innopac	City University, London	k.A.	Nein
Innopac	Glasgow University	k.A.	Nein
PICA	GBV Gemeinsamer Verbundkatalog	21,200.000	**Ja**
PICA	HeBIS Verbundkatalog	10,600.000	**Ja**
PICA	Deutsche Bibliothek, Frankfurt	4,700.000	Nein
PICA	Deutsche Bibliothek, Leipzig	5,900.000	Nein
PICA	TIB/UB Hannover	k.A.	Nein
PICA	Technische Universität Ilmenau	500.000	**Ja**
Sirsi	London School of Economics	2,000.000	**Ja**
Sirsi	Imperial College, London	500.000	**Ja**
Sirsi	Brunel University, London	300.000	**Ja**
SISIS	Universität Köln (USB)	k.A.	Nein
SISIS	Deutsche Zentralbibl. f. Medizin, Köln	k.A.	Nein
Talis	London Metropolitan University	k.A.	Nein
Talis	Manchester Metropolitan University	1,000.000	Nein
Talis	Birmingham University	k.A.	Nein
Voyager	Library of Congress	12,000.000	Nein
Voyager	Cambridge University	k.A.	**Ja**
Voyager	University of Wales Aberystwyth	500.000	**Ja**
k.A.	Schwedischer Verbundkatalog (LIBRIS)	5,000.000	Nein

Tabelle 2: Verfügbarkeit von Relevance Ranking in ausgewählten weiteren OPACs

Als Beispiele für sehr grosse OPACs mit Relevance Ranking können die Verbundkataloge des GBV und von HEBIS genannt werden (PICA-Systeme).

Anzeige Gewichte	Default Sortierung	Einstellbar durch User	Sonstiges
-	-	-	
-	-	-	
-	-	-	
-	-	-	
Nein	**Ja**	Ja	
-	-	-	
-	**Ja**	Nein	nur bei Titel-Suche!
-	-	-	
-	-	-	
-	-	-	
Nein	Nein	Ja	
Nein	Nein	Ja	
-	-	-	
-	-	-	
-	-	-	
Nein	Nein	Ja	
Nein	Nein	Ja	nur bis 200 Treffer
Symbol	**Ja**	Ja	nur bis 200 Treffer
Symbol	Nein	Ja	nur bis 200 Treffer
-	-	-	
-	-	-	
-	-	-	
-	-	-	
-	-	-	
-	-	-	
Symbol	**Ja**	Ja	nur bei „Alle Felder“-Suche und Boolescher Suche
Symbol	**Ja**	Ja	nur bei Keyword-Suche
-	-	-	

In beiden Katalogen kann der Benutzer (vor dem Absetzen der Suche) selbst entscheiden, ob das Rechercheergebnis nach Erscheinungsjahr oder

Relevanz sortiert werden soll. Gleiches gilt aber auch für kleinere PICA-OPACs wie etwa das Lokalsystem der Technischen Universität Ilmenau (Abb.1). In allen PICA-Katalogen wurde überdies eine besonders gute Systemperformance beim Ranking (aber auch bei anderen Sortierfunktionen) registriert.

Sortierung der Trefferlisten: Im Suchmenü finden Sie das Menü, in dem Sie auswählen können, ob Ihre Rechercheergebnisse nach Relevanz oder Erscheinungsjahr sortiert werden sollen.

Erscheinungsjahr: die aktuellsten Titel werden am Anfang der Liste angezeigt.

Relevanz: Der Vorteil der Sortierung nach Relevanz liegt darin, dass diejenigen Treffer in der Ergebnismenge zuerst angezeigt werden, die Ihrem Suchanliegen vermutlich am meisten entsprechen. Je weiter unten ein Treffer auf Trefferliste steht, um so geringer ist seine Gewichtung. Nach dem Auffinden aller Dokumente, die der Suchanfrage entsprechen, erfolgt eine inhaltliche Analyse nach folgenden Kriterien:

Ähnlichkeit: Je grösser die Ähnlichkeit zwischen dem vorgegebenen Suchbegriff und dem gefundenen im Dokument ist, um so höher ist die Gewichtung/Relevanz des Dokuments

Häufigkeit: Wie oft kommt der Suchbegriff in einem Dokument vor?

Länge: Wie lang ist ein Dokument, verglichen mit anderen? Kurze Texte werden in ihrer Gewichtung / Relevanz höher eingestuft als lange.

Reziproke Dokumenthäufigkeit: Seltene Suchbegriffe erhalten ein grösseres Gewicht

Umfang der Übereinstimmung: Suchbegriffe, die für ein Dokument charakteristisch sind, erhalten ein grösseres Gewicht

Abb. 1: Auszüge aus der Online-Hilfe des OPACs (PICA) der Technischen Universität Ilmenau

In dem besonders in Grossbritannien verbreiteten System SIRSI kann Relevance Ranking sowohl durch die Bibliothek (Imperial College) als auch durch die Benutzer als Sortierkriterium voreingestellt werden. Die Gewichte werden in der Ergebnisliste durch die Anzeige eines farbigen „Balkens" unterschiedlicher Länge symbolisiert. Kurioserweise ist jedoch in den von uns untersuchten SIRSI-OPACs Relevance Ranking ausschliesslich für Treffermengen von max. 200 Dokumenten möglich! Hilfetexte bzw. sonstige Hinweise zum Ranking konnten in diesen OPACs nicht gefunden werden. Auch das vor allem im angelsächsischen Raum populäre System VOYAGER

bietet Relevance Ranking an. Zwar fanden wir dieses Feature im Katalog der Library of Congress nicht implementiert, wohl jedoch in den OPACs der Universitäten Cambridge und Wales-Aberystwyth. Das Ranking erfolgt relativ rasch; die Gewichte werden in der Ergebnisliste bildlich dargestellt (unterschiedliche Anzahl rechteckiger bzw. runder Symbole). Bei beiden Katalogen ist Ranking die systemseitig voreingestellte Sortieroption! Dies gilt jedoch nicht für alle Arten von Suchen, sondern nur für die „Alle Felder"-Suche (Cambridge und Wales-Aberystwyth) bzw. die Boole'sche Suche (nur Cambridge). Die Benutzerinformation erscheint adäquat (Abb. 2).

Search Tips:

A **Keyword Anywhere** search finds words, phrases or names anywhere in a catalogue record. This type of search has features similar to those of an Internet search engine (e.g. Google, AltaVista). The result of this search option is a list of items ordered by relevance.

A **Boolean** search is a more flexible keyword search and finds words, phrases or names anywhere in a catalogue record. Searches can be as simple as a single word or include arrangements of multiple terms or phrases that can be restricted to specific fields of a catalog record. Boolean operators (AND, OR, NOT) may be used to combine search terms in order to narrow or broaden a search.
The result of this search option is a list of items ordered by relevance.

The factors that affect the **relevance ranking** are:

- Uniqueness of search terms within the database
- Proximity of search terms to each other within the bibliographic record
- Proportion of search terms present in a bibliographic record
- Fields (subject heading, author, title) in which the search terms occur. Some of the fields, such as subject, carry extra weight.

Abb. 2: Auszüge aus der Online-Hilfe des „Newton"-OPACs (Voyager) der Universität Cambridge

Auch im britischen Verbundkatalog COPAC ist Relevance Ranking als Default-Sortieroption für einen bestimmten Suchtyp eingerichtet, in diesem Fall für die (reine, unverknüpfte) Titelstichwortsuche. Da es sich bei der gereihten Treffermenge um das Produkt einer systemintern vermengten Phrasen- und Stichwortsuche handelt, kann eine solche Ergebnisliste nicht mehr umsortiert werden.

Schliesslich führte unsere Recherche zu dem v.a. in Australien und Neuseeland eingesetzten Bibliothekssystem C2 der Firma CONTEC. Dieses System verwendet keine Boole'sche Suche, sondern basiert offensichtlich auf einem probabilistischen Retrieval, demzufolge die Treffer nach Relevanz sortiert werden (Default-Einstellung).

Eine Übersicht über die hier besprochenen OPACs mit Relevance Ranking sowie deren Web-Adressen zeigt Tabelle 3.

Institution / OPAC	**Adresse (URL)**
Österreichischer Verbundkatalog	http://opac.bibvb.ac.at/acc01
Verbund für Bildung und Kultur	http://opac.bibvb.ac.at/vbk01
Vorarlberger Landesbibliothek	http://vlb-katalog.vorarlberg.at/F/-/?local_base=vlb01
GBV Gemeinsamer Verbundkatalog	http://www.gbv.de/cgi-bin/wwwobn2psi?DB=2.1&LNG=DU
HeBIS Verbundkatalog	http://webcbs.rz.uni-frankfurt.de/
Technische Universität Ilmenau	http://katalog.bibliothek.tu-ilmenau.de/start/
COPAC (UK)	http://www.copac.ac.uk/copac/
University College London	http://library.ucl.ac.uk
Imperial College, London	http://www.imperial.ac.uk/library/resources/cataccess.htm
Brunel University, London	http://library.brunel.ac.uk:8080/uhtbin/webcat/
London School of Economics	http://catalogue.lse.ac.uk
Cambridge University	http://ul-newton.lib.cam.ac.uk
University of Wales Aberystwyth	http://voyager.aber.ac.uk/
Université de Paris III	http://maestro.scd.univ-paris3.fr/
Univ. of California Santa Barbara	http://pegasus.library.ucsb.edu/aleph
Central Queensland Institute of TAFE	http://cqitopac.tafe.net/

Tabelle 3: Web-Adressen von OPACs mit Relevance Ranking [überprüft: 13.11.2003]

Relevance Ranking im Aleph 500 OPAC

In diesem Abschnitt soll der (vorläufige) Wissensstand der Autoren über die Details der Möglichkeiten präsentiert werden, die ein OPAC unter Aleph 500 im Hinblick auf Relevance Ranking zu bieten hat.

In der zur Verfügung stehenden Aleph-Dokumentation wird dieser Aspekt leider nur kursorisch abgehandelt (Ex Libris, 2001, 41f). Immerhin geht daraus hervor, dass für die Berechnung der Gewichte eine gebräuchliche Formel (Termhäufigkeit x inverse Datenbankhäufigkeit) verwendet werde und dass darüberhinaus die Möglichkeit bestehe, in dieser Formel auch noch Gewichte („location weights") für einzelne Indizes zu berücksichtigen (Standardwert = 0001 in Spalte 9 der Indexdefinitionstabelle „tab00.lng"). Um dieses relativ spärliche Informationsgerüst zu erweitern, versuchten wir, durch „Experimentieren" in einem kontrollierten, kleinen Testdatenbestand den oben angeführten Algorithmus für die Relevanzberechnung zu untersuchen bzw. im Detail zu ergründen.

(1) Zunächst sollte durch die systematische Manipulation von Testdatensätzen in Erfahrung gebracht werden, welche Parameter tatsächlich in die Relevanzberechnung einfliessen.

- Handelt es sich bei der „Termhäufigkeit" um die *einfache Termfrequenz* (Häufigkeit Term je Dokument) oder um die *relative Termfrequenz* (einfache Termfrequenz / Gesamtzahl der Terme im Dokument)? Hiezu konnte klargestellt werden, dass der Aleph-Algorithmus die einfache Termfrequenz verwendet und die Länge des Dokumentes nicht berücksichtigt.
- Wird tatsächlich die „Datenbankhäufigkeit" (Häufigkeit des Terms in der gesamten Datenbank) oder nicht doch – wie dies vermutlich sinnvoller wäre – die *Dokumenthäufigkeit* (Frequenz der Dokumente je Term) verwendet? Dazu konnten wir nachweisen, dass sich die Anzahl des Indexterms in der Gesamtdatenbank *nicht*, die Anzahl der Dokumente zum Indexterm in der Gesamtdatenbank hingegen *sehr wohl* auf die Relevanzberechnung auswirkt. Dies bedeutet, dass die o.a. Aleph-Dokumentation, in der explizit von der Datenbankhäufigkeit die Rede ist, in diesem Punkt irrt.

(2) Des weiteren interessierten uns die Möglichkeiten, die sich durch die Anwendung der oben erwähnten „location weights" bieten könnten. Diese Gewichte sind ausschliesslich durch den Systembibliothekar und nicht etwa benutzerseitig einstellbar. Sie beziehen sich nur auf einen jeweiligen *Index* (z.B. Alle Felder, Titelstichwort, Schlagwort) und nicht auf einzelne *Kategorien* (z.B. Hauptsachtitel, Zusätze zum Hauptsachtitel – hier wären etwa unterschiedliche Gewichte sinnvoll). Nur die letzten beiden Stellen des eigentlich vierstelligen Parameters dürfen verwendet werden (01 bis 99).

- Wenn man einzelne Indizes mit unterschiedlichen Gewichten ausstattet (z.B. Titelstichwortindex mit Gewicht 99, Schlagwortindex mit Gewicht 01), so werden beim Ranking tatsächlich deutliche Unterschiede gegenüber der Standardeinstellung (01 bei allen Indizes) erkennbar. Allerdings basiert der dabei zur Anwendung kommende Algorithmus offensichtlich *nicht* bloss auf einer einfachen Multiplikation, da wir die vom System für das Ranking berechneten Gewichte mittels einer solchen Rechenoperation nicht nachzuvollziehen vermochten.
- Unklar war ausserdem, ob das System auch bei einer „Alle Felder"-Suche die vordefinierten Gewichte entsprechend den Indizes, zu denen die jeweiligen Suchterme grundsätzlich gehören, oder aber das ebenfalls in der Indexdefinitionstabelle angeführte Gewicht für den Alle-Felder-Index für die Berechnung des Ranking heranzieht (so gehört beispielsweise ein Stichwort aus dem Hauptsachtitel primär zum Index „Titelstichwörter" und erst in zweiter Linie zum Index „Alle Felder"). Hiezu konnte festgestellt werden, dass Aleph auch bei der Recherche im Index „Alle Felder" bei der Relevanzberechnung eine Zuordnung der Indexterme zu den „primären" Indizes vornimmt. Dies ist insofern von Bedeutung, als der überwiegende Anteil aller Publikumsrecherchen im Index „Alle Felder" erfolgt.
- Ob die Verwendung vordefinierter Gewichte für einzelne Indizes überhaupt einen Verbesserungseffekt auf das Ranking besitzt bzw. welche konkreten Gewichtungsfaktoren am besten für einzelne Indizes verwendet werden sollten, konnte im Rahmen dieser kurzen Betrachtung nicht überprüft werden.

(3) Da im System Aleph 500 die Möglichkeit besteht, zu den Schlagwörtern der in den MAB-Kategorien 902, 907, ..., 947 enthaltenen RSWK-Ketten auch die Verweisungsformen aus der SWD einzuspielen, weisen die entsprechenden Indizes (Schlagwortindex bzw. Index „Alle Felder") auch diese Terme auf, was sogar als *rudimentäre* Form einer automatischen Indexierung bezeichnet werden könnte. Dadurch führt bspw. die Eingabe des Nondeskriptors „Bibliotheksautomation" zu Treffern, die mit der SWD-Vorzugsbenennung „Bibliothek/Automation" beschlagwortet sind. Da dieser Mechanismus in den OPACs des Österreichischen Bibliothekenverbundes genutzt wird, hatten wir ursprünglich beabsichtigt, auch die Auswirkungen dieser Indexanreicherung durch die Normdateien im Hinblick auf das Relevance Ranking zu untersuchen. Da dies jedoch in der vorhandenen Testumgebung ohne grösseren Aufwand nicht realisierbar ist, wurde diese Frage vorläufig zurückgestellt.

Probleme unter Aleph 500

(1) Dokumentation
Seit der Einführung der Version 14.2 (Frühjahr 2001) wurde wiederholt – jedoch ergebnislos – versucht, von der Herstellerfirma nähere Informationen zur Funktion „Ergebnisliste gewichten" zu erhalten, da die in der Dokumentation (s.o.) gebotenen Informationen zu knapp erscheinen. Insbesondere fehlen jegliche Hinweise auf sinnvolle Einstellungen für die „location weights"; hiezu müssten eigentlich Erfahrungen der Entwickler bzw. sogar systematische Untersuchungen vorliegen. Ausserdem fehlen Erläuterungen zu den Auswirkungen des Normdaten-Expands auf die Gewichtung und Hinweise darauf, wie dies vom Systembibliothekar beeinflusst werden kann.

(2) Relevance Ranking nicht als Sortierkriterium voreinstellbar
Die Anzeige einer gewichteten Ergebnisliste sollte als Default voreinstellbar bzw. vom Benutzer bei der Eingabe der Suchbegriffe auswählbar sein. Bislang ist es nur möglich, eine bereits ermittelte Treffermenge („Ergebnisliste") durch Betätigen eines Buttons gewichten zu lassen. Besonders für grosse OPACs (z.B. Verbundkatalog) wäre dies eine sinnvolle Parametrisierungs-Option.

(3) Ausgabeprobleme: Anzeige der Gewichte
In Information-Retrieval-Systemen ist es üblich, Relevanzgewichtungen in Prozenten anzuzeigen. Beim Aleph-OPAC erfolgt jedoch eine Promille-Anzeige (1000 statt 100,0%). Im Verbundkatalog wurde dies von uns durch ein einfaches, jedoch immer noch unbefriedigendes Script verändert (z.B. wird statt 981 nun 98.1 angezeigt, statt 710 jedoch 71; jeweils eine Kommastelle und rechtsbündige Darstellung wären erwünscht). Dies sollte verbessert bzw. parametrisierbar gemacht werden.

(4) Ausgabeprobleme: Sortierfehler
Beim Blättern in einer gewichteten Ergebnisliste kommt es im Verbundkatalog häufig, wenngleich nicht immer, zu einer falschen Sortierung. Folgende Fälle wurden bisher registriert: 1) falsche Sortierfolge (d.h. nicht nach dem Gewicht) bereits ab der zweiten Seite; 2) zunächst beim Vorblättern richtige, beim Zurückblättern aber plötzlich falsche Sortierfolge; 3) nach der Rückkehr aus einer Vollanzeige falsche Sortierfolge der Ergebnisliste.

(5) Performance

Die Funktion „Ergebnisliste gewichten“ ist durch eine absolut unzureichende Performance gekennzeichnet. Gerade bei grossen Treffermengen versagt die Funktion oft vollständig. So führt z.B. der Versuch, ca. 4.000 Treffer zu gewichten, manchmal zum Abbruch (Timeout), manchmal schafft das System die Gewichtung nach etwa einer Minute(!) Wirklich grosse Treffermengen (z.B. 60.000) konnten bisher nicht gewichtet werden (immer Timeout). Selbst der Versuch, nur 430 Titel zu gewichten, scheiterte manchmal (sofern es funktionierte, waren 30 Sekunden das Minimum). Man vergleiche dagegen die Leistung des GBV-OPACs (s.o.), wo gerade bei grossen Treffermengen die Gewichtung vom System selbst vorgeschlagen wird, z.B.: „Ihre Eingabe war: (...) Es sind 612.912 Treffer (Tipp: Stellen Sie die Sortierung auf Relevanz ein).“ Selbst bei dieser enormen Trefferzahl wies der GBV-OPAC eine auffallend gute Performance auf (Gewichtung unter 2 Sekunden!)

Perspektiven

Aufgrund der Charakteristika des Ranking-Features im Aleph-OPAC sowie der bisherigen Erfahrungen damit kann man sich des Eindrucks nicht erwehren, dass das Thema „Relevance Ranking“ für die Herstellerfirma keine grosse Bedeutung besitzt. Möglicherweise könnte jedoch im Wege der internationalen Anwendergruppe „ICAU“ eine Verbesserung der derzeitigen Situation erreicht werden.

Dies wäre besonders im Hinblick auf eine zukünftig mögliche Anreicherung der bibliographischen Datensätze (automatische Indexierung, Inhaltsverzeichnisse, Abstracts) wünschenswert, da unter dieser Perspektive Relevance Ranking sicherlich noch notwendiger und nützlicher wäre als bisher.

Dennoch müsste angesichts der mangelnden Parametrisierungs-Optionen und der unzufriedenstellenden Systemperformance darüber diskutiert werden, ob Relevance Ranking im Verbundkatalog gegenwärtig weiter angeboten oder bis zu einer zukünftigen Verbesserung dieses Features ausser Betrieb genommen werden sollte.

Eine solche Entscheidung sollte allerdings unter Miteinbeziehung des Benutzerverhaltens getroffen werden, d.h. es wäre zu prüfen, ob Logfile-Analysen hinsichtlich der Inanspruchnahme des Ranking-Features durchführbar sind und was diese ggfs. erbringen.

Darüberhinaus wäre es gewiss vorteilhaft, den hier beschriebenen Informationsstand noch zu erweitern, vor allem hinsichtlich folgender Aspekte:

- Indexanreicherung durch Normdateien
- vordefinierte Gewichte („location weights“)
- Option der Hinzunahme weiterer Terme in den Gewichtungsprozess

Die Autoren beabsichtigen, dies durch weitere Überprüfungen des Algorithmus anhand eines kleinen, exakt definierten und kontrollierten Datenpools im kommenden Jahr durchzuführen.

Dr. Otto Oberhauser & Ing. Josef Labner
Die Österreichische Bibliothekenverbund- und Service GmbH
Garnisongasse 7/21, A 1090 Wien.
Tel.: +43 1 4035158-17; +43 1 4035158-44
E-Mail: otto.oberhauser@bibvb.ac.at, josef.labner@bibvb.ac.at

Literatur

Ex Libris (2001). *Database Management Guide: Version 14.2, Part 1: UTIL A–H.*

Gödert, W. (1996). Vom OPAC zum Hyperkatalog: Suchen und Navigieren. In: Hermes, H.-J.; Wätjen, H.-J. (Hrsg.) *Erschliessen, Suchen, Finden: Vorträge aus den bibliothekarischen Arbeitsgruppen der 19. und 20. Jahrestagungen (Basel 1995 / Freiburg 1996) der Gesellschaft für Klassifikation.* Oldenburg. Bibliotheks- u. Informationssystem d. Univ. 75–90.

Google. (2002). *Warum man Google benutzen sollte.* URL: http://www.google.at/why_use.html [besucht: 04.11.2003]

Güde, B.; Götsch, D. (2002). *Suchen im Internet – Literaturrecherche.* Proseminar Internetwerkzeuge – Wintersemester 2002/2003. Skript zum Vortrag 25.11.2002. URL: http://vsis-www.informatik.uni-hamburg.de/teaching/ws-02.03/prosint/termin06/vortrag_prosint.pdf [besucht: 31.10.2003]

Lepsky, K. (1996). Vom OPAC zum Hyperkatalog: Daten und Indexierung. In: Hermes, H.-J.; Wätjen, H.-J. (Hrsg.) *Erschliessen, Suchen, Finden: Vorträge aus den bibliothekarischen Arbeitsgruppen der 19. und 20. Jahrestagungen (Basel 1995 / Freiburg 1996) der Gesellschaft für Klassifikation.* Oldenburg. Bibliotheks- u. Informationssystem d. Univ. 65–73.

Rowley, J. E.; Farrow, J. (2000). *Organizing Knowledge: An Introduction to Managing Access to Information*. 3rd ed. Aldershot: Gower.

Salton, G.; McGill, M. J. (1987). *Information Retrieval: Grundlegendes für Informationswissenschaftler.* Hamburg: McGraw-Hill. [Engl. Orig. 1983]

Stets, P. (1999). AW: OPAC-Umfrage: Ergebnisse. *InetBib* [online]. 22.06.1999. URL: http://www.ub.uni-dortmund.de/listen/inetbib/msg12288.html [besucht: 3.11.2003]